Title:

Noninvariant Measurement in Rater-mediated Assessments of Teaching Quality

Authors and Affiliations:

Ben Kelcey
University of Cincinnati
ben.kelcey@gmail.com

Background / Context:

Valid and reliable measurement of teaching is essential to evaluating and improving teacher effectiveness and advancing large-scale policy-relevant research in education (Raudenbush & Sadoff, 2008). One increasingly common component of teaching evaluations is the direct observation of teachers in their classrooms. Classroom observations have been long viewed as a promising way to evaluate and develop teachers because they anchor assessments in specific and observable criteria (Gitomer, 2009).

Despite the potential of classroom observations to identify strengths and address specific weaknesses in teachers' practices, a significant problem with observed teaching scores is that they confound construct-irrelevant variation with persistent teaching quality (i.e., observed scores are not independent of the characteristics of a specific observation). When measuring persistent differences among teachers in terms of their quality, idiosyncratic features of these observations (e.g., atypical lesson, rater effects, etc.) can introduce substantial construct-irrelevant variance or measurement error. For example, the MET study (2012) reported that across each of the instruments, construct-irrelevant variation constituted as much as 60% to 90% of the variation in observed scores. Left untreated, construct-irrelevant variance has the potential to unfairly affect outcomes and undermine the reliability and validity of classroom observations (Messick, 1989).

In observational assessments of teaching, one key source of construct-irrelevant variation is the differences among raters. Research has shown that even after extensive training there are important differences in how raters interpret evidence and that these differences potentially introduce variability in the structure of the scale established by the guiding rubric/instrument (e.g., Eckes, 2009; Hill, Charlambous, & Kraft, 2012; Engelhard, 2002). For example, some raters may perceive the difference between two adjacent ratings in terms of the implied quality levels to be much farther apart than other raters do (Eckes, 2009). This type of variability violates assumptions of measurement invariance and precludes meaningful comparisons of scores assigned by different raters because scores are no longer on a common scale and have a consistent meaning across raters. When we obtain a low rating for a teacher, the extent to which that low rating reflects the ineffectiveness of the teacher or the relative severity of the assigned rater(s) is confounded. As a result, the extent to which we can place teachers on a common scale depends heavily on the extent to which raters share a common basis for interpreting, evaluating, and discriminating among different levels of teaching quality.

Purpose / Objective / Research Question / Focus of Study:

The focus of this study was to develop and investigate a set of psychometric methods that accommodate, as best as possible, measurement noninvariance among raters. Our approach draws on multilevel item response theory and extends it by introducing rater-specific item parameters for noninvariant items. To introduce rater-specific item parameters, we structured item parameter differences among raters as random deflections from overall item parameters. Our approach relaxes assumptions of measurement invariance across raters by explicitly adjusting for the different ways in which raters use instruments to establish an approximately invariant scale.

To empirically investigate the value of the method, we applied it to the measurement of persistent teaching quality using classroom observations. In turn, we examined two questions: (1) to what extent is measurement noninvariant across raters in classroom observations and (2) to

what extent does adopting an approximate measurement invariance approach improve the fit and predictive validity of the observation scores over simple averages?

Setting & Participants:

This study draws on a subsample of the Measures of Effective Teaching data (e.g., MET, 2012). We used 650 teachers who had valid classroom observations and valid 2011 value-added scores on the standardized mathematics achievement test. Teachers and their classrooms were drawn from six urban districts across the United States and in this subsample each teacher was observed, on average, on three different lessons.

Intervention / Program / Practice:

In this study, we investigated the value of our new method using two different observations systems. The first observation system we investigated was the Framework for Teaching (Danielson, 2011). In the MET study, the Framework for Teaching instrument described teaching through two different primary domains: classroom environment and instruction. Our investigation and analyses in this study focused on the domain of instruction. This domain was evaluated by raters using four indicators: communicating with students, using questions and discussion techniques, engaging students in learning, and using assessment in instruction. The second observation system we investigated was the Mathematical Quality of Instruction lite (Hill et al., 2008). In this study, our analyses focus on four indicators of quality: errors and imprecision, student participation in meaning making and reasoning, richness, and working with students and mathematics.

Significance / Novelty of study:

The results reported in the MET (2012) and other studies suggest that even with extensive training there is still significant variation among raters in terms of their judgments. Despite this variability, current and forthcoming policy initiatives are likely to include classroom observation scores to evaluate individual teachers. Although classroom observations will likely only be a single piece of information used in these evaluations, they will frequently play a key role. To this end, developing measurement models that are more tightly attuned to the types of measurement errors present in classroom observation systems is likely to improve the comparability of scores across irrelevant facets (e.g., raters) and the validity of and the precision with which we can describe teaching quality through classroom observations. In turn, the increased comparability, validity, and precision are likely to strengthen our understanding and the legitimacy of using classroom observations to evaluate teachers.

Statistical, Measurement, or Econometric Model:

To estimate persistent teaching quality using rater-mediated classroom observations, we developed a cross-classified multilevel random item effects graded response model. Our approach first drew on conventional item response theory models so that observed indicator scores are treated as fallible ordinal ratings stemming from a latent trait of persistent teaching quality (Hambleton, Swaminathan, & Rogers, 1991). Second, because each teacher was measured across multiple observations rated by different raters, we leveraged (cross-classified) multilevel item response theory to introduce random effects for observations (Fox, 2010). Third, to provide a basis for placing teachers rated by different raters on a similar scale, we relaxed assumptions of measurement invariance across raters to account for rater differences by allowing

indicator properties (i.e., discrimination and threshold) to vary across raters. By allowing indicator properties to vary across raters, random item effects approximate measurement invariance by compensating for the reality that raters may have complex differences in their scales (e.g., not just simple shifts in severity but varying thresholds and abilities to discriminate among levels). Let

$$P(Y_{iotr} = k) = \Phi(a_i \theta_t + a_{ir}\alpha_{ot} - d_{ir}^{k-1}) - \Phi(a_i \theta_t + a_{ir}\alpha_{otr} - d_{ir}^{k}) \quad (1)$$

Here Φ is the normal cumulative distribution function, Y_{iotr} is the score for item *i* in observation *o* for teacher *t* rated by rater *r,* θ_t represents teacher *t*'s persistent level of quality, a_{ir} is the discrimination parameter for item *i* rated by rater *r* and each a_{ir} has a distribution across raters such that $a_{ir} \sim N(a_i, \sigma_{a,i}^2)$. Further α_{otr} is the quality deviation specific to observation *o* for teacher *t,* let *K* represent the number of categories items are graded on with *k* as a specific category and let $d_i^{(1)}, ..., d_i^{(K-1)}$ be a set of *K*-1 ordered item difficulty thresholds where each d_{ir}^k has a distribution across raters such that $d_{ir}^k \sim N(d_i^k, \sigma_{d,i}^2)$. To identify the scale, let $\alpha \sim N(0,1)$ and $\theta \sim N(0, \sigma_T^2)$. Scores and parameter estimates were obtained using Bayesian estimation and noninformative priors (Asparouhov & Muthen, 2012; Gelman, Carlin, Stern & Rubin, 2004).

By introducing random effects for item parameters (across raters), the rater-specific item discriminations, a_{ir}, and thresholds, d_{ir}^k capture and adjust for differences among raters in how they use items. As a result, $\sigma_{a,i}^2$ and $\sigma_{d,i}^2$ describes the variability of these effects across all raters and the extent to which there is measurement noninvariance across raters.

Findings / Results:

Although there are several important differences among the models in terms of invariance, fit, and the tenability of assumptions, we highlight two aspects of the results. First, we investigated the extent to which the calibration of each instrument suggested items were invariant across raters. Our results indicated that there was significant variability in the item parameters across raters suggesting both metric and scalar noninvariance (see Table 2). If raters' uses of indicators were invariant across raters, the variance of the item parameters (displayed in Table 1) across raters would be zero. Evident from Table 1, each of the estimated variance components was different from zero (and in most cases significantly different) indicating that observations rated by different raters were not on a common scale.

We next investigated how teacher-specific estimates of persistent teaching quality predicted value-added estimates of teacher effectiveness and compared these estimates to simple averages. The results are presented in Table 2. Similar to the results found in the MET report (2012), the relationship between teachers' averaged Mathematical Quality of Instruction classroom observation ratings and their value-added scores was fairly low and insignificant (Table 2). However, by adjusting for construct-irrelevant variance and measurement noninvariance among raters, the relationship nearly tripled and became statistically significant. We saw similar but smaller differences when applying the method to observations scored with the Framework for Teaching instrument. The relationship between teachers' Framework for Teaching instruction scores and their value-added scores was approximately 25% larger when using the item response theory based method versus simple averages (Table 2).

Conclusions:

The comparability of scores assigned by different raters is a well-known and complex problem because raters may vary in how they interpret and score observations. The results suggested that such measurement noninvariance is also present in raters' application of teaching quality rubrics to classroom observations. For both the Mathematical Quality of Instruction and Framework for Teaching instruments, our results indicated that item parameters varied across raters thus suggesting that a common scale of teaching quality is not preserved across observations rated by different raters. A practical consequence of this rater nonequivalence was the attenuation of the relationships between different measures of teaching. For both of the observation instruments examined, we saw that the relationships between observation scores and value-added scores were improved upon by adjusting for measurement noninvariance across raters. The results underscored the practical importance of establishing a common scale across raters. However, the differing magnitudes of the improvement across instruments also suggested that there may be uneven benefits from the proposed methods across instruments. We saw significant gains for the Mathematical Quality of Instruction instrument in terms of how it related to value-added estimates but we saw much smaller gains for the same relationship with the Framework for Teaching instrument. Because instruments have different visions of teaching quality and use different systems and sets of competencies to operationalize these theories, instruments may vary in their sensitivity to rater effects. Instruments that are less sensitive to rater differences may reduce the value of the proposed method.

Evidence from this study suggests the promise of random item effect models to address measurement non-invariance in rater-mediated assessments. However, there is a question of whether random item effects and the associated approximate measurement invariance can adequately compensate for differences in the scales raters use. The potential value of this method needs to be carefully studied to understand the extent to which random item effect models to can effectively address non-invariant conditions.

Appendices

Appendix A. References

Asparouhov, T. & Muthen, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts and parameters. http://www.statmodel.com/download/NCME12.pdf

Danielson, C. (December 2010/January 2011). Evaluations that help teachers learn. The Effective Educator, 4, 35-39.

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium* (pp. 43–73). Frankfurt, Germany: Lang

Fox, J. (2010). Bayesian Item Response Modeling: Theory and Applications. Springer.
Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). Bayesian Data Analysis (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Gitomer, D. (2009). Measurement Issues and Assessment for Teaching Quality. London: Sage Publications.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and application. Boston, MA: Kluwer-Nijhoff Publishing.

Hill, H., Charlambous, C., & Kraft, M. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. Educational Researcher, 41, 2, 56-64.

Hill, H., Blunk, M., Charalambous, C., Lewis, J., Phelps, G., Sleep. L., & Ball, D. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An Exploratory Study. *Cognition and Instruction*, 26, pp. 430-511. (Messick, 1989)

MET Measures of Effective Teaching (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Retrieved from the Bill and Melinda Gates Foundation Measures of Effective Teaching website: metproject.org

Raudenbush, S., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness,* 1, 138-154.

Appendix B. Tables and Figures

Table 1: Variance in item parameters across raters

	Thresholds	Discriminations
Mathematical Quality of Instruction		
Errors and imprecision	0.33	0.06
Student participation in meaning making and reasoning	0.80	0.16
Richness	0.27	0.13
Working with students and mathematics	0.74	0.24
Framework for Teaching		
Using questions and discussion techniques	1.36	0.23
Communicating with students	1.92	0.17
Engaging students in learning	1.66	0.15
Using assessment in instruction	1.86	0.27

Table 2: Relation between value-added and classroom observation scores

	Standardized Regression Coefficient	*t*-value
Mathematical Quality of Instruction		
Averages	0.05(0.04)	1.32
Item response theory	0.13(0.05)	2.76
Framework for Teaching		
Averages	0.14(0.04)	3.86
Item response theory	0.17(0.04)	4.27